A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data

L Wynants[a,b], W Bouwmeester[c,d], KGM Moons[c], M Moerbeek[e], D Timmerman[f,g], S Van Huffel[a,b], B Van Calster[f,h], Y Vergouwe[h]

[a] KU Leuven Department of Electrical Engineering-ESAT, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, box 2446, 3001 Leuven, Belgium. Laure.wynants@esat.kuleuven.be, Sabine.VanHuffel@esat.kuleuven.be.

[b] KU Leuven iMinds Medical IT Department, Kasteelpark Arenberg 10, box 2446, 3001 Leuven, Belgium.

[c] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, P.O. Box 85500, 3508 GA Utrecht, The Netherlands. K.G.M.Moons@umcutrecht.nl.

[d] Pharmerit B.V., Marten Meesweg 107, 3068 AV, Rotterdam, The Netherlands. bouwmeester.w@kpnmail.nl.

[e] Department of Methodology and Statistics, Utrecht University, PO Box 80140 3508 TC Utrecht. m.moerbeek@uu.nl.

[f] KU Leuven Department of Development and Regeneration, Herestraat 49 box 7003, 3000 Leuven, Belgium. dirk.timmerman@uzleuven.be, ben.vancalster@med.kuleuven.be.

[g]Department of Obstetrics and Gynaecology, University Hospitals Leuven, Leuven, Belgium. dirk.timmerman@uzleuven.be.

[h] Center for Medical Decision Sciences, Department of Public Health, Erasmus Medical Center, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands. y.vergouwe@erasmusmc.nl.

Corresponding author: Laure Wynants, ESAT - STADIUS Centre for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10 box 2446, 3001 Leuven, Belgium. +32 16 3 21065. Laure.wynants@esat.kuleuven.be.

This article has been accepted for publication in Journal of Clinical Epidemiology Published by Oxford University Press. DOI: http://dx.doi.org/10.1016/j.jclinepi.2015.02.002.

Abstract

Objective: The study aims to investigate the influence of the amount of clustering (intraclass correlation [ICC]=0%, 5%, or 20%), the number of events per variable or candidate predictor (EPV=5, 10, 20, or 50), and backward variable selection on the performance of prediction models.

Study Design and Setting: Researchers frequently combine data from several centers to develop clinical prediction models. In our simulation study, we developed models from clustered training data using multilevel logistic regression and validated them in external data.

Results: The amount of clustering was not meaningfully associated with the models' predictive performance. The median calibration slope of models built in samples with EPV=5 and strong clustering (ICC=20%) was 0.71. With EPV=5 and ICC=0%, it was 0.72. A higher EPVrelated to an increased performance: the calibration slope was 0.85 at EPV=10 and ICC=20% and 0.96 at EPV=50 and ICC=20%. Variable selection sometimes led to a substantial relative bias in the estimated predictor effects (up to 118% at EPV=5), but this had little influence on the model's performance in our simulations.

Conclusion: We recommend at least ten EPV to fit prediction models in clustered data using logistic regression. Up to fifty EPV may be needed when variable selection is performed.

Keywords: clustered data, multicenter study, events per variable, logistic model, prediction model, simulation study

Running Head Title: Events per variable in clustered data

Word count: 3760

What is new?

- The number of events per variable (EPV) can be used to guide sample size decisions in clustered data. There are no existing guidelines on the required number of events relative to the number of (candidate) predictors under study when data is clustered. We recommend to have at least 10 EPV for predefined models, although up to 50 may be needed when performing variable selection.
- Unlike previous studies, this study does not only investigate the influence of sample size and clustering on the bias in regression estimates, but also on the predictive performance of the regression model.
- This study further illustrates that besides the number of EPV, also the total number of observations contributes to the accuracy of regression coefficients and the performance of the prediction model.

1 Introduction

Clinical prediction models are useful aids to making a diagnosis or prognosis. They are often constructed using multivariable logistic regression if the health outcome of interest is binary [1]. Researchers proposed to use a sample including at least ten events per variable or candidate predictor (EPV) for the development of a prediction model [2, 3]. The number of events is the number of observations in the smallest outcome category of the binary outcome. The number of variables, henceforward referred to as predictors, should be interpreted more broadly as the number of parameters to be considered. E.g., more than one parameter per predictor must be estimated when polynomial terms are used to model a non-linear effect or when dummy coding is used to model the effect of a qualitative predictor with more than two categories. Hence, in a

dataset of 500 observations with 50 events in total, only five parameters should be estimated in order to obtain an EPV of ten. Some researchers have proposed upward [4, 5] and downward [6] adjustments for the EPV guideline, stating that the required EPV is influenced by the size of the predictor effects, the correlations among predictors, the prevalence of dichotomous predictors, and the predictor selection strategy. Predictor selection in particular may result in strongly biased estimated regression coefficients in small samples, which decreases the predictive performance of the prediction model when used in individuals other than those from which the model was developed [4].

Multicenter consortia are gaining popularity: recruiting patients from different sites produces a representative sample and reduces recruitment times [7, 8]. They yield clustered datasets, i.e. datasets with dependent observations, since patients from one center may have more in common than patients from different centers [9]. These datasets can be analyzed using multilevel regression (also known as mixed or random effects regression, or hierarchical modeling), in order to build a prediction model [1, 10, 11]. Multilevel regression enables the incorporation of center-specific intercepts and predictor effects [12]. Simulation studies have shown that the amount of clustering, the number of clusters, and cluster size influence the accuracy and precision of the parameter estimates, especially the random effect variances and the predictor effects at the cluster level [13-15]. These studies, however, considered neither predictive performance nor the required number of EPV.

Here, we use simulated data and an empirical example of the classification of ovarian tumors to study the effect of the number of EPV for prediction modeling with clustered data when using multilevel logistic regression. We hypothesize that the number of EPV will influence both parameter estimates and the performance of the developed prediction model. The effective

sample size [12] is smaller than the number of participants, because they have not been sampled independently. The amount of clustering may therefore have a negative effect on the model's performance. However, this impact may be limited, since only the estimated predictor effects are used in prediction, and previous research has shown that these are usually estimated with limited bias [13-15].

2 Design of the Simulation Study

We studied the influence of the number of EPV on the parameter estimates and the predictive performance of the logistic multilevel regression model. We have built models in samples with varying numbers of EPV. These samples were drawn from a source population with a certain degree of clustering. The predictive performance of the models was tested in the source population. In what follows, we describe the design of the simulation study (technical details in Web Appendix 1, R code in Web Appendix 2) and the performance measures we evaluated.

2.1 The source populations

We created source populations (of approximately 100,000 observations across 200 clusters) with an ICC of 0%, 5%, or 20%. These values were chosen to reflect situations without clustering, with moderate clustering, and with extreme but realistic clustering in multicenter prediction research [16]. The source populations were generated according to a model with four uncorrelated continuous predictors (X_1~N(0,1), X_2~N(0,0.6), X_3~N(0,0.4), X_4~N(0,0.2)), four uncorrelated dichotomous predictors (X_5 to X_8, with prevalence 0.2, 0.3, 0.3, and 0.4 respectively), and a random intercept of which the variance was determined by the ICC. All regression coefficients were set at 0.8 to achieve a level of discrimination that is common in the applied literature. The overall intercept was set at -2.1 to obtain an outcome event rate of 0.3,

which is common for health outcomes in prediction research. For one population (with ICC=20%), we introduced a correlation between the random intercepts and X_1 and X_5, such that predictors were unequally distributed across clusters. The mean of X_1 ranged from -1.13 in the cluster with the lowest random intercept to 1.44 in the cluster with the highest random intercept, while the prevalence of X_5 ranged from 12% to 29%.

2.2 Sampling

We sampled datasets from the source populations. The number of EPV was set at 5, 10, 20, or 50. These values reflect popular choices and guidelines for prediction research. In clustered data, the EPV is determined as $\text{EPV} = N \times p/k$, where N is the sum of all individual cluster sizes n_j (j=1 to J), i.e. the total sample size, p is the sample's event rate, and k is the number of parameters in the model to be estimated, including the random intercept variance. We defined several simulation conditions, grouped into sets of simulations which are characterized by the parameters that are varied (Table 1): the average number of observations per cluster ($\bar{n}_j$=5, 10, 20, 30, or 50) in set 1 to 3 and set 6, the number of clusters (J=5, 10, 20, 30, or 50) in set 4, and the sample's event rate (p=0.05, 0.1, 0.2, 0.3, or 0.5) in set 5. The event rate in the sample determines the total sample size when EPV and the number of predictors are fixed, since $N = \text{EPV} \times k/p$. We drew 500 datasets per simulation condition.

Insert Table 1 here.

2.3 Model building

Random intercept models were fitted in each sample. They were either predefined models using all eight true predictors (k=9, sets 1, 4, 5, 6), full models including the eight true predictors and eight noise variables (k=17, set 2), or reduced models determined through backward variable selection (α=0.1) starting from all sixteen candidate predictors (k=17, set 3) (Table 1). The noise variables had the same distributions as the true predictors. To limit the amount of simulations, the presence of noise predictors and variable selection (sets 2 and 3), the effects of the ICC (set 1), and the correlations between random intercepts and predictors (set 6) were only studied in samples with J=30 and p=0.3.

We used the following criteria for model convergence: 10 to 100 iterations to fit the model, a change of less than 10^{-5} in deviances of the models fitted in the last two iterations, and no outlying estimated regression coefficients and standard errors (visual inspection). All models fulfilled these criteria and no samples needed to be deleted.

2.4 Model evaluation

All models were evaluated in terms of the accuracy of estimated regression parameters and the predictive performance.

2.4.1 Bias in the estimated regression coefficients

We compared each estimated regression coefficient $\widehat{\beta}$ to β_{sp}, the regression coefficient when the model was built in the source population. The percentage of relative bias in the estimated regression coefficients was defined as $100 \times (\hat{\beta} - \beta_{sp})/\beta_{sp}$. The use of β_{sp} ensures that random error originating from generating a source population is not included in the computation of the bias in the estimated regression coefficients. β_{sp} was between 0.770 and 0.857 for all predictors

in all source populations. The relative bias of the estimated random intercept variance was computed analogously.

2.4.2Predictive performance

The predictive performance of the resulting prediction models was tested in the corresponding source population. We used predictions for the average center, omitting the random intercept, to be able to make predictions in clusters that were not represented in the sample [11, 17].

The validated C-index or concordance probability (C) [18] measured the discriminatory performance of the developed model. It was obtained by testing the fitted model in the source population. The calibration slope (b) [1, 19] was used to evaluate the accuracy of predicted probabilities in the source population. It is obtained through logistic regression of the event indicator against the linear predictor of the prediction model. If b is smaller than one, there is overfitting, i.e., the predicted probabilities are too extreme (too close to zero or one); if b is larger than one, there is underfitting, i.e., the predicted probabilities are not extreme enough.

We also computed the within-cluster C-index (C_{within}) and the within-cluster calibration slope (b_{within}) to evaluate the performance at the cluster level instead of the population level. The former is a weighted combination of center-specific C-indices [20], the latter is estimated using logistic regression with random cluster intercepts and a random cluster calibration slope [10].

The obtained C-indices and calibration slopes were compared to those of a model developed and evaluated in the source population (henceforward C_{sp} and b_{sp}), which serve as upper limits for discriminatory power and calibration in the given population. The relative C-index and the relative calibration slope were computed as $100 \times C/C_{sp}$ and $100 \times b/b_{sp}$ (or $100 \times C_{within}/C_{sp\ \ within}$ and $100 \times b_{within}/b_{sp\ within}$ for the within-cluster measures). Note that b_{sp} will deviate from one because predictions for the average center, omitting the random intercepts, are

used. Even if random intercepts are used in prediction, the calibration slope will deviate from one, because the cluster-specific random intercepts are shrunk to zero [12].

All simulations and calculations were performed in R version 2.14.0 [21]. The lmer function from the lme4 package was used to fit multilevel logistic regression models using Laplace approximation [22] and the rms package was used for model evaluation [18].

3 Results of the Simulation Study

3.1 Data clustering and the number of events per variable

The amount of clustering(ICC) did not influence the relative bias of the estimated regression coefficients (Figure 1A, representing results from simulation set 1). The median bias was close to 0% at each ICC. The bias of the estimated regression coefficients related to the EPV: the interquartile range (IQR) of the relative bias was largest for the lowest EPV values. At ICC=20%, the median relative bias of $\hat{\beta}_4$ was -10% at EPV=5 and -2% at EPV=50. Similar patterns were observed for the other regression coefficients (Table A.1). The random intercept variances were often underestimated, but a large number of EPV benefited estimation (median relative bias -15.2% at EPV=5 to -5.2% at EPV=50 for ICC=20%).

Model performance also related to the EPV, and minimally to the ICC. The within-cluster C-index of the model fitted and evaluated in the source population was 0.78 in each source population (ICC=0%, 5% and 20%). The relative within-cluster discrimination of models fitted in the samples was the lowest for the samples with EPV=5, with median values of around 97% (Figure 1B). The calibration slope of the models fitted in the source populations was 1.00 at each ICC and the relative performance of the models fitted in the samples were similar for varying

ICCs (Figure 1C). At ICC=20%, the median relative within-cluster calibration slope increased from 71.6% at EPV=5 to 96.6% at EPV=50. The IQR of the calibration slope also decreased with increasing EPV.

The same patterns were observed for the relative overall C-index and the relative overall calibration slope (Figure A.1A and A.1B), but the overall C-indices and calibration slopes of the models fitted in the source populations did decrease with increasing ICC (C_{sp} 0.78, 0.78, and 0.76, and b_{sp} 1.00, 0.97, and 0.88 for ICC=0%, 5%, and 20% respectively).

Insert Figure 1 here.

3.2 Variable selection

A high EPV was required to ensure the inclusion of important predictors and to prevent bias in the estimates when using backward variable selection (α=0.1). At EPV=5, X_4 was selected in only 24% percent of the samples. To ensure the selection of X_4 in >90% of the samples, 50 EPV were required (Table 2). The median relative bias of $\hat{\beta}_4$ (conditional on selection) was 118% at EPV=5 and disappeared at EPV=50 (0.05% relative bias) (Figure 2A, representing results from simulation conditions 1.3, 1.6, 1.9 and 1.12, and sets 2-3). The selection bias for other predictors was negligible. Predefined models that included all candidate predictors showed biases similar to predefined models that included only the eight true predictors for the regression coefficients of the eight true predictors. The smaller IQRs in models with sixteen variables reflect the larger sample sizes required to obtain the related EPV values with sixteen rather than eight predictors.

Models containing all sixteen candidate predictors had a slightly better median relative within-cluster discrimination and a lower median relative within-cluster calibration slope than

models after variable selection at EPV=5 (95.9% versus 95.3% of $C_{sp\ within}$=0.78, and 67.3% vs 72.1% of $b_{sp\ within}$=1.00, Figure 2B and 2C). This is in accordance with earlier findings [23]. The differences in performance reduced as EPV increased. At EPV=10, 20 and 50, the predictive performance of the models after variable selection was comparable to the model including only true predictors. The lower relative within-cluster C-index after variable selection at EPV=5 may be explained by the frequent exclusion of relevant predictors (Table 2). The same pattern was observed for the overall C-index and calibration slope (Figure A.1C and A.1D).

Insert Figure 2 here.

Insert Table 2 here.

3.3 Sample size

The IQR of the relative bias in the estimated regression coefficients and the model performance related to the total sample size. Samples with EPV=5 and a total sample size of 900 showed a smaller range of relative bias than samples with EPV=5 and a total sample size of 150 (Figure 3A, representing results of simulation conditions 1.3, 1.6, 1.9, 1.12, and sets 4-5). A large total sample size and a large number of clusters reduced the relative bias in the estimated random intercept variance (Figure A.2). The relative within-cluster C-indices and calibration slopes of models fitted in the larger samples were higher (Figure 3B and 3C). For a given total sample size and number of EPV, samples with many small clusters yielded a predictive performance comparable to samples with a few large clusters. Note that at EPV=5, the IQRs of the bias in the estimated regression coefficients, the within-cluster C-indices and the within-cluster calibration

slopes were slightly smaller in samples with a few large clusters. The overall performance measures showed similar patterns (Figure A.1E and A.1F).

Insert Figure 3 here

3.4 Random cluster effects correlated with predictors

When the assumption of independence between random intercepts and predictors was violated, $\hat{\beta}_1$ was positively biased (FigureA. 3A, representing the results of simulation conditions 1.3, 1.6, 1.9, 1.12, and set 6). The same holds for $\hat{\beta}_5$. These regression coefficients accounted for the cluster-level association of X_1 and X_5 with the random intercept, yielding more severely underestimated random intercept variances (Figure A.3B). Since $\hat{\beta}_1$ and $\hat{\beta}_5$ contributed to explaining differences between clusters, the relative overall C-index was increased compared to the situation in which predictors and random intercepts were independent (Figure A.3D). This could not be observed for the within-cluster C-index, as $\hat{\beta}_1$ and $\hat{\beta}_5$ did not enable a better discrimination within clusters (Figure A.3C). Finally, because $\hat{\beta}_1$ and $\hat{\beta}_5$ were positively biased, overfitting was more problematic (Figure A.3E and A.3F).

4 Empirical Example

We developed clinical prediction models to diagnose ovarian cancer using data from the International Ovarian Tumor Analysis (IOTA) group [24, 25]. We analyzed clinical and ultrasound information on 5912 patients with ovarian masses from 24 hospitals, collected between 1999 and 2012. The data collected up until 2005 (n=1571, 9 centers, 409 (26%) malignant tumors) was used for the development of prediction models for tumor malignancy that

included random intercepts for hospitals. We drew one hundred samples of 409 events and 1162 non-events from the training set, with replacement. We considered seven predictors (Table 3). Together with the random intercept variance, this yielded eight parameters to estimate (EPV=51). We further drew one hundred random subsets of 154 patients (40 malignancies) to obtain development samples with EPV=5. The regression coefficients of the prediction models are shown in Table 3. The most extreme estimates were obtained when variable selection was performed with EPV=5. Note that the random intercept variances estimated by the models fitted with EPV=5 were lower than the estimates obtained with EPV=51. This is in line with the findings of the simulation study, where the random intercept variance was underestimated in samples with EPV=5 (figure A.2).

Insert Table 3 here.

The data collected after 2005 (n=4341, 22 centers, 1522 (35%) malignancies) was used for the models' validation. The models' performance depended on the number of EPV. The full model fitted with EPV=51 gave a median validated within-cluster calibration slope of 0.966 (IQR 0.934 to 1.011) while the full model fitted with EPV=5 gave a much lower median calibration slope of 0.763 (IQR 0.509 to 0.855). The median validated within-cluster C-indices were 0.862 (IQR 0.860 to 0.864) and 0.854 (IQR 0.844 to 0.859) respectively. The effects of EPV were similar when backward variable selection (α=0.10) was used. The median within-cluster calibration slopes were 0.970 (IQR 0.939 to 1.015) and 0.815 (0.735 to 0.933) for the models fitted with EPV=51 and EPV=5 respectively. The median within-cluster C-indices were 0.862 (IQR 0.861 to 0.863) and 0.851 (IQR 0.842 to 0.859) respectively. Compared to a model developed with a

high EPV (EPV=51), a low EPV (EPV=5) resulted in a similar discriminative ability of the prediction model, but more overfitting (Table 4).

Insert Table 4 here.

5 Discussion

Our simulation research showed that the number of EPV determines the bias in parameter estimates of prediction models developed in clustered data using multilevel logistic regression, as well as the resulting models' predictive performance (discrimination and calibration) in external data. At EPV=5 and ICC=20%, predictions were too extreme, but this overfitting disappeared at EPV=50. Models built on samples with EPV=50 were also slightly better at discriminating between events and non-events. This was illustrated in our case study. The amount of clustering was not meaningfully associated with the models' predictive performance. Our simulation results also suggest that larger samples provide better models for a given EPV. This means that non-events contribute to the stability of the prediction model, provided the number of events is sufficient. When variable selection was performed, a high number of EPV was needed to ensure the inclusion of relevant predictors and reduce estimation bias in the estimated predictor effects.

In accordance with earlier proposals for non-clustered data [2, 3], we recommend to use at least ten EPV to fit a predefined prediction model in clustered data, although up to fifty EPV may be needed when stepwise variable selection is applied [4]. There were no negative effects of clustering in the simulation conditions we have considered if the number of EPVwas sufficiently large. However, it must be noted that it is impossible to obtain an optimal overall calibration

slope for a random intercept model, if predictions for the average center (omitting random intercept estimates) are used. Even if the model was fitted on the data of the entire source population, the median overall calibration slope was 0.88 rather than 1 when the ICC was 20%. The within-cluster calibration slope does not suffer from this issue.

Common formulas for power calculations for the design of experiments take into account the ICC, because clustering reduces the effective sample size and necessitates larger samples [12]. Nonetheless, the estimation of the regression coefficients used for prediction purposes is little influenced by clustering [13-15], provided that the assumption of independence between predictors and random effects is not violated, and there is no interaction between predictor effects and cluster [9]. Hence, the existing EPV guidelines apply in clustered data if a random cluster intercept is added to the prediction model and the estimation of this additional parameter is accounted for in the EPV calculation.

Multilevel models are very useful tools when analyzing clustered datasets. Random slopes can be used to investigate interactions between predictor effects and cluster, and random intercepts can be used to model differences in outcome prevalence across clusters. It is difficult to reliably estimate the variance of the random intercept, but our results show that estimation improves when the number of clusters increases. This confirms earlier findings [13-15]. It has been recommended to collect data in at least fifty clusters [14], although random effect variances may still be slightly underestimated with hundreds of clusters [15]. In reality, however, the number of centers in multicenter research is most often smaller than fifty, and is determined by weighing benefits, such as the reduction of recruitment times, against practical concerns, to maintain the manageability of the study. Our findings also demonstrate that a high EPV benefits

the estimation of the ICC, regardless of the total sample size or the number of clusters. This has not been studied previously.

We believe that the focus on the predictive performance of models, alongside the estimation of regression parameters, is a strength of this study, because the purpose of clinical prediction modeling is to make reliable predictions for new subjects [1]. For the same reason, other evaluation criteria such as confidence interval coverage, type I error rates, and the power of statistical tests are of lesser importance. The uncertainty inherent in model building was acknowledged by studying variable selection. We have used within-cluster performance measures to acknowledge the use of prediction models in separate centers [26]. Our simulation study, like all simulation studies, is restricted by our choice of simulation parameter settings. However, we have chosen practically relevant settings for our simulation parameters, such as the ICC, the number of EPV, and the number of clusters. Furthermore, we did not assume equal numbers of observations in clusters, as this hardly ever occurs in multicenter research [8]. The parameters that were not varied, such as the estimation method (Laplace approximation, [12]) and the variable selection method (backward variable selection [1, 18]), were set at generally accepted or recommended choices. A final strength of our study is that we studied the effect of violating the assumption of independence between predictors and random effects [12]. Because two predictors were dependent on the random intercepts, they were also correlated with each other. We did not consider random predictor effects, assuming homogeneity of predictor effects across clusters. Center-level predictors were not included in the study. Reliable estimation of center-level effects will depend on the number of clusters.

In conclusion, this study acknowledges the clustered nature of datasets collected in multicenter research, and shows that the number of events per variable is useful in guiding sample size decisions when the aim is to develop prediction models using clustered data.

Acknowledgements

This work was supported by a PhD fellowship from the Flanders' Agency for Innovation by Science and Technology (IWT Vlaanderen) to LW; the Research Foundation-Flanders (FWO) (a travel grant to LW, ; a fundamental clinical research fellowship to DT, and project grant G049312N), the Netherlands Organization for Scientific Research (grant 917.11.383 to YV, projects 9120.8004 and 918.10.615 to KM) and ZonMw (grant 17088.25029 to WB). This work was further supported by the Research Council KUL [GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC)], the Flemish Government [iMinds Medical Information Technologies SBO 2014, FWO (grant G049312N)] and the Belgian Federal Science Policy Office [IUAP P7/19/ (DYSCO, 'Dynamical systems, control and optimization', 2012-2017)]. The funding sources had no involvement in the study design, the data collection, analysis or interpretation, the writing of the article or the decision to submit the article for publication.

The authors declare that they have no competing interests.

References

1. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York, NY: Springer US; 2009.

2. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373-1379. DOI: 10.1016/S0895-4356(96)00236-3.

3. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361-387. DOI: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

4. Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol. 1999;52(10):935-942. DOI: 10.1016/S0895-4356(99)00103-1.

5. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. J Clin Epidemiol. 2011;64(9):993-1000. DOI: 10.1016/j.jclinepi.2010.11.012.

6. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol. 2007;165(6):710-718. DOI: 10.1093/aje/kwk052.

7. Sprague S, Matta JM, Bhandari M, Dodgin D, Clark CR, Kregor P, et al. Multicenter collaboration in observational research: improving generalizability and efficiency. J Bone Joint Surg Am. 2009;91 Suppl 3:80-86. DOI: 10.2106/jbjs.h.01623.

8. Senn S. Some controversies in planning and analysing multi-centre trials. Stat Med. 1998;17(15-16):1753-1765; discussion 1799-1800. DOI: 10.1002/(SICI)1097-0258(19980815/30)17:15/16<1753::AID-SIM977>3.0.CO;2-X.

9. Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for center in multicenter studies: an overview. Ann Intern Med. 2001;135(2):112-123. DOI: 10.7326/0003-4819-135-2-200107170-00012.

10. Bouwmeester W, Twisk J, Kappen T, Klei W, Moons K, Vergouwe Y. Prediction models for clustered data: comparison of a random intercept and standard regression model. BMC Med Res Methodol. 2013;13(1). DOI: 10.1186/1471-2288-13-19.

11. Debray TPA, Moons KGM, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat Med. 2013;32(18):3158-3180. DOI: 10.1002/sim.5732.

12. Snijders TAB, Bosker RJ. Multilevel analysis : an introduction to basic and advanced multilevel modeling. 2nd ed. Snijders TAB, Bosker RJ, editors. London: Sage; 2012.

13. Maas CJM, Hox JJ. Sufficient Sample Sizes for Multilevel Modeling. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences. 2005;1(3):86-92. DOI: 10.1027/1614-2241.1.3.86

14. Moineddin R, Matheson FI, Glazier RH. A simulation study of sample size for multilevel logistic regression models. BMC Med Res Methodol. 2007;7(34). DOI: 10.1186/1471-2288-7-34.

15. Paccagnella O. Sample Size and Accuracy of Estimates in Multilevel Models. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences. 2011;7(3):111-120. DOI: 10.1027/1614-2241/a000029.

16. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. J Clin Epidemiol. 2004;57(8):785-794. DOI: 10.1016/j.jclinepi.2003.12.013.

17. Skrondal A, Rabe-Hesketh S. Prediction in multilevel generalized linear models. Journal of the Royal Statistical Society Series A (Statistics in Society). 2009;172(3):659-687. DOI: 10.1111/j.1467-985X.2009.00587.x.

18. Harrell FE. Regression modeling strategies : with applications to linear models, logistic regression, and survival analysis. Harrell FE, Jr., editor. New York, NY: Springer; 2001.

19. Cox DR. Two Further Applications of a Model for Binary Regression. Biometrika. 1958;45(3/4):562-565. DOI: 10.2307/2333203.

20. Van Oirbeek R, Lesaffre E. Assessing the predictive ability of a multilevel binary regression model. Computational Statistics & Data Analysis. 2012;56(6):1966-1980. DOI: 10.1016/j.csda.2011.11.023.

21. R Development Core Team. R: A language and environment for statistical computing Vienna, Austria: R Foundation for Statistical Computing; 2011 [October 7, 2014]. Available from: http://www.R-project.org/.

22. Bates D, Maechler M, Bolker B. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-42. 2011 [October 7, 2014]. Available from: http://CRAN.R-project.org/package=lme4.

23. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med. 2000;19(8):1059-1079. DOI: 10.1002/(SICI)1097-0258(20000430)19:8<1059::AID-SIM412>3.0.CO;2-0.

24. Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) group. Ultrasound in Obstetrics and Gynecology. 2000;16(5):500-505. DOI: 10.1046/j.1469-0705.2000.00287.x.

25. Kaijser J, Bourne T, Valentin L, Sayasneh A, Van Holsbeke C, Vergote I, et al. Improving strategies for diagnosing ovarian cancer: a summary of the International Ovarian Tumor Analysis (IOTA) studies. Ultrasound Obstet Gynecol. 2013;41(1):9. DOI: 10.1002/uog.12323.

26. van Klaveren D, Steyerberg E, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. BMC Med Res Methodol. 2014;14(1). DOI: 10.1186/1471-2288-14-5.

Tables

Table 1. Simulation Conditions

Set	Condition	Population		Sample				Model		EPV
		ICC (%)	Corr (X,u_j)	J	$\bar{n}_j$	N	p	k	Backward selection	
1	1.1	**0**	0	30	**5**	**150**	0.3	8+1	No	**5**
1	1.2	**5**	0	30	**5**	**150**	0.3	8+1	No	**5**
1	1.3	**20**	0	30	**5**	**150**	0.3	8+1	No	**5**
1	1.4	**0**	0	30	**10**	**300**	0.3	8+1	No	**10**
1	1.5	**5**	0	30	**10**	**300**	0.3	8+1	No	**10**
1	1.6	**20**	0	30	**10**	**300**	0.3	8+1	No	**10**
1	1.7	**0**	0	30	**20**	**600**	0.3	8+1	No	**20**
1	1.8	**5**	0	30	**20**	**600**	0.3	8+1	No	**20**
1	1.9	**20**	0	30	**20**	**600**	0.3	8+1	No	**20**
1	1.10	**0**	0	30	**50**	**1500**	0.3	8+1	No	**50**
1	1.11	**5**	0	30	**50**	**1500**	0.3	8+1	No	**50**
1	1.12	**20**	0	30	**50**	**1500**	0.3	8+1	No	**50**
2	2.1	20	0	30	**9**	**270**	0.3	8+8 noise+1	No	**5**
2	2.2	20	0	30	**18**	**540**	0.3	8+8 noise+1	No	**10**
2	2.3	20	0	30	**36**	**1080**	0.3	8+8 noise+1	No	**20**
2	2.4	20	0	30	**89**	**2670**	0.3	8+8 noise+1	No	**50**
3	3.1	20	0	30	**9**	**270**	0.3	8+8 noise+1	Yes	**5**
3	3.2	20	0	30	**18**	**540**	0.3	8+8 noise+1	Yes	**10**
3	3.3	20	0	30	**36**	**1080**	0.3	8+8 noise+1	Yes	**20**
3	3.4	20	0	30	**89**	**2670**	0.3	8+8 noise+1	Yes	**50**
4	4.1	20	0	**5**	30	**150**	0.3	8+1	No	**5**
4	4.2	20	0	**10**	30	**300**	0.3	8+1	No	**10**
4	4.3	20	0	**20**	30	**600**	0.3	8+1	No	**20**
4	4.4	20	0	**50**	30	**1500**	0.3	8+1	No	**50**
5	5.1	20	0	30	30	900	**0.05**	8+1	No	**5**
5	5.2	20	0	30	30	900	**0.1**	8+1	No	**10**
5	5.3	20	0	30	30	900	**0.2**	8+1	No	**20**
5	5.4	20	0	30	30	900	**0.5**	8+1	No	**50**
6	6.1	20	>0	30	**5**	**150**	0.3	8+1	No	**5**
6	6.2	20	>0	30	**10**	**300**	0.3	8+1	No	**10**
6	6.3	20	>0	30	**20**	**600**	0.3	8+1	No	**20**
6	6.4	20	>0	30	**50**	**1500**	0.3	8+1	No	**50**

[a] The varying parameters within each set of conditions are indicated in bold

Legend: ICC: intraclass correlation, corr(X, u_j): correlation between predictors and the random intercept, J: the number of clusters, $\bar{n}_j$: the average number of observations per cluster, N: the total number of observations, p: the event rate of the outcome in the sample, k: the number of parameters to be estimated, including the random intercept, EPV: the number of events per variable.

Table 2. The Selection Frequency of Predictors

	EPV=5	**EPV=10**	**EPV=20**	**EPV=50**
X_1	100%	100%	100%	100%
X_2	88%	99%	100%	100%
X_3	60%	85%	99%	100%
X_4	24%	43%	65%	93%
X_5	64%	87%	100%	100%
X_6	72%	95%	100%	100%
X_7	76%	95%	100%	100%
X_8	79%	95%	100%	100%

Legend: EPV: events per variable

Table 3. The Fitted Models for Ovarian Tumor Diagnosis in 100 bootstrap samples, based on EPV=5 versus EPV=51, without (Model 1 and Model 2) and with (Model 3 and Model 4) Backward Variable Selection (α=0.10)

Predictor	Model 1 (EPV=51) $Median\ (IQR)$	Model 2 (EPV=5) $Median\ (IQR)$	Model 3 (EPV=51) $Median\ (IQR)$ $[selection\ frequency]$	Model 4 (EPV=5) $Median\ (IQR)$ $[selection\ frequency]$
Age (per 10 years)	0.38 (0.35 to 0.42)	0.40 (0.27 to 0.55)	0.39 (0.36 to 0.42) [100]	0.48 (0.39 to 0.61) [76]
Maximum diameter of the lesion (per 10 mm)	0.12 (0.11 to 0.13)	0.13 (0.08 to 0.18)	0.12 (0.11 to 0.13) [100]	0.16 (0.11 to 0.19) [82]
Presence of solid tissue in the lesion (yes versus no)	3.25 (3.13 to 3.40)	3.69 (2.98 to 4.25)	3.23 (3.08 to 3.39) [100]	3.49 (2.91 to 4.08) [91]
Family history of ovarian cancer (yes versus no)	0.40 (0.20 to 61)	0.40 (-0.08 to 1.01)	0.69 (0.62 to 0.82) [30]	2.17 (1.99 to 2.26) [4]
Current use of hormonal therapy (yes versus no)	-0.35 (-0.46 to -0.24)	-0.47 (-0.87 to -0.07)	-0.43 (-0.54 to 0.38) [60]	-1.24 (-1.39 to -1.15) [15].
Pelvic pain during examination (yes versus no)	-0.13 (-0.25 to 0.01)	-0.32 (-0.62 to 0.03)	-0.39 (-0.43 to -0.33) [21]	-1.35 (-1.63 to -1.23) [8]
Presence of papillary structures (yes versus no)	-0.07 (-0.20 to 0.08)	-0.15 (-0.57 to 0.28)	-0.28 (-0.32 to -0.26) [14]	1.03 (-1.14 to 1.17) [24]
Intercept	-6.50 (-6.80 to -6.25)	-7.20 (-8.49 to –5.79)	-6.56 (-6.81 to -6.28)	-6.68 (-7.85 to -5.04)
Random intercept variance	0.59 (0.45 to 0.75)	0.20 (0.00 to 0.67)	00.61 (0.45 to 0.75)	0.19 (0.00 to 0.67)

Legend: EPV: events per variable, n.s.: not significant and not retained during variable selection

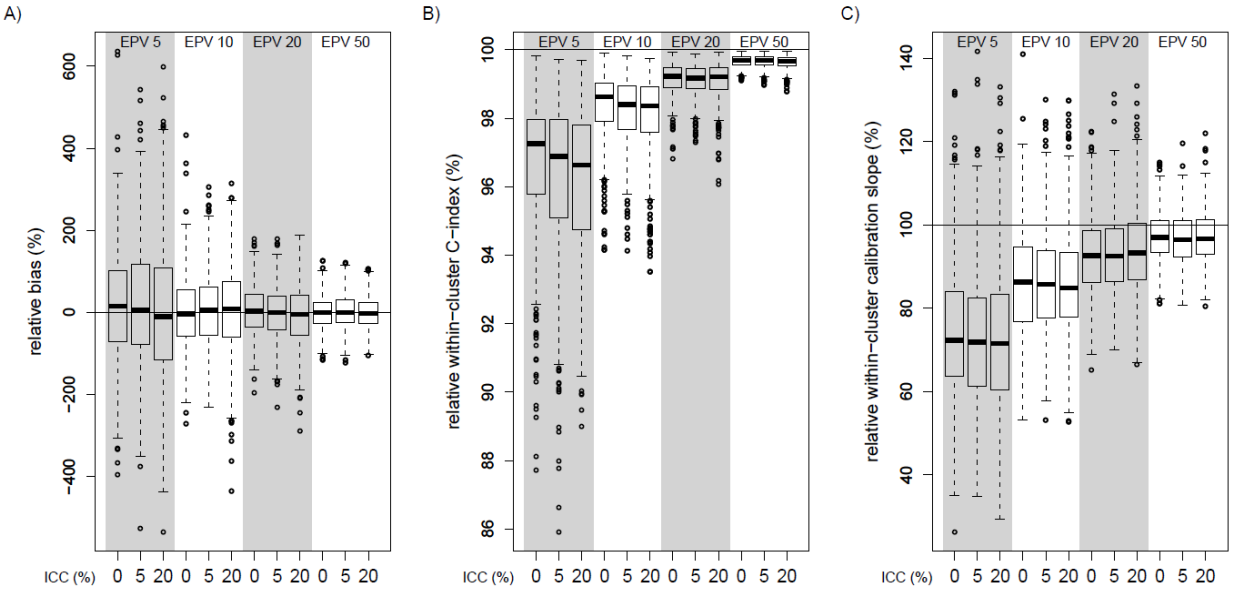
Table 4. The Performance of the Fitted Models for Ovarian Tumor Diagnosis, based on EPV=5 versus EPV=51 in 100 bootstrap samples, without (Model 1 and Model 2) and with (Model 3 and Model 4) Backward Variable Selection (α=0.10)

	Model 1 (EPV=51)	**Model 2 (EPV=5)**	**Model 3 (EPV=51)**	**Model 4 (EPV=5)**
Within-cluster C-index (IQR)	0.862 (0.860 to 0.864)	0.854 (0.844 to 0.859)	0.862 (0.861 to 0.863)	0.851 (0.842 to 0.859)
Within-cluster calibration slope (IQR)	0.966 (0.934 to 1.011)	0.763 (0.509 to 0.855)	0.970 (0.939 to 1.015)	0.815 (0.735 to 0.933)
Overall C-index (IQR)	0.882 (0.881 to 0.884)	0.873 (0.865 to 0.878)	0.882 (0.881 to 0.883)	0.868 (0.858 to 0.878)
Overall calibration slope (IQR)	0.955 (0.923 to 1.002)	0.751 (0.477 to 0.850)	0.959 (0.925 to 1.00)	0.809 (0.732 to 0.925)

Legend: EPV: events per variable, n.s.: not significant and not retained during variable selection

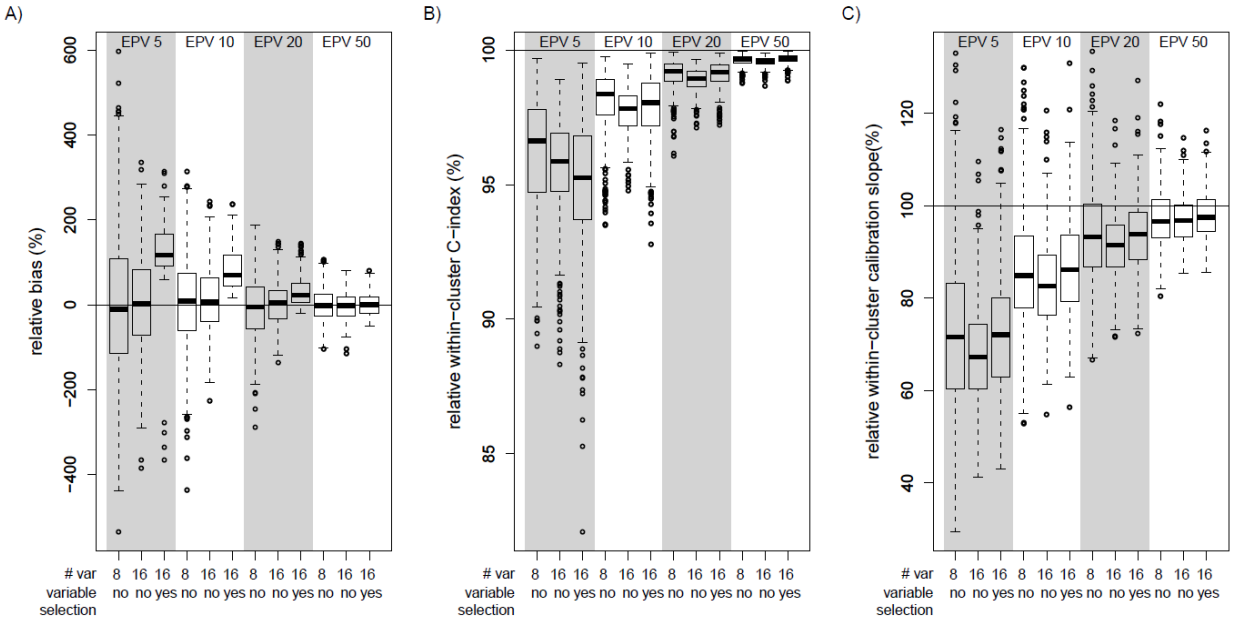
Figures

Figure 1. The Relative Bias in Estimated Regression Coefficients and the Predictive Performance in Relation to the Number of Events per Variable (EPV) and the Amount of Clustering (ICC)



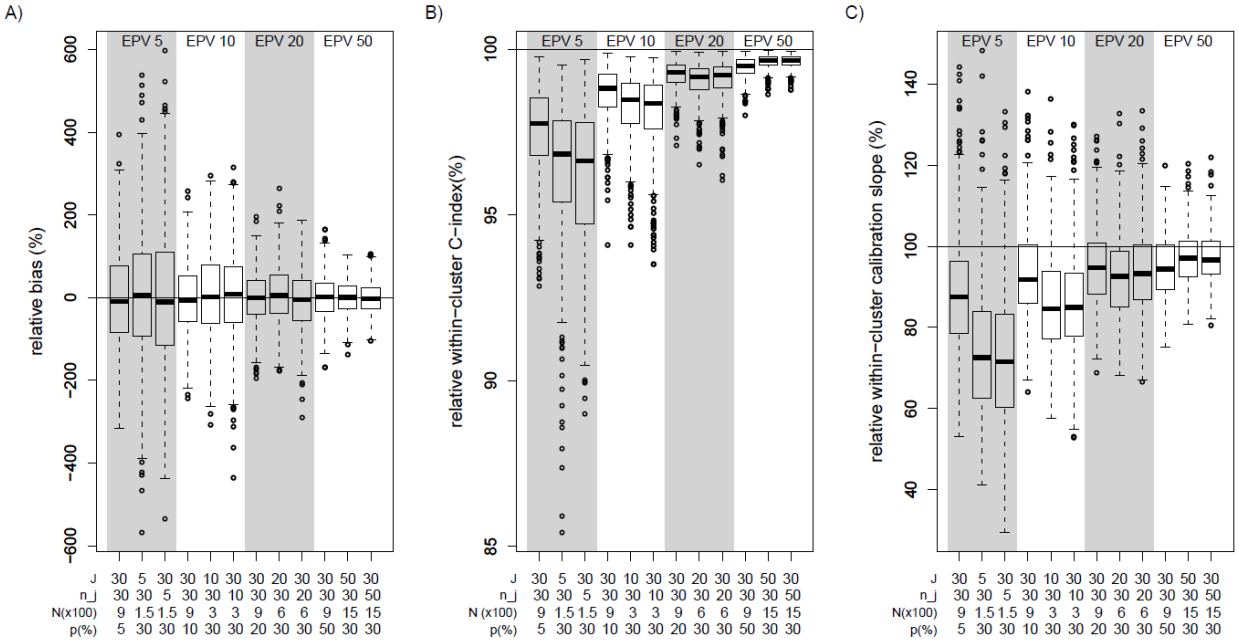
Legend: A) the relative bias (%) in the estimated regression coefficient $\hat{\beta}_4$; B) the relative within-cluster discrimination (%), computed as $100 \times (C_{within}/C_{sp\ within})$, where $C_{sp\ within}$ is the within-cluster C-index of a model fitted and evaluated in the source population; C) the relative within-cluster calibration slope (%), defined as $100 \times (b_{within}/b_{sp\ within})$, where $b_{sp\ within}$ is the within-cluster calibration slope of a model fitted and evaluated in the source population. The box indicates the interquartile range (IQR), the fat horizontal line within the box indicates the median. Whiskers extend to the lowest and highest data points still within 1.5 IQR of the box. Outliers beyond these points are represented by dots.

Figure 2. The Relative Bias in Estimated Regression Coefficients and the Predictive Performance in Relation to the Number of Events per Variable (EPV) and Backward Variable Selection (ICC=20%)



Legend: #var: number of candidate predictors. A) the relative bias (%) in the estimated regression coefficient $\hat{\beta}_4$; B) the relative within-cluster discrimination (%), computed as $100 \times (C_{within}/C_{sp\ within})$, where $C_{sp\ within}$ is the within-cluster C-index of a model fitted and evaluated in the source population; C) the relative within-cluster calibration slope (%), defined as $100 \times (b_{within}/b_{sp\ within})$, where $b_{sp\ within}$ is the within-cluster calibration slope of a model fitted and evaluated in the source population. The box indicates the interquartile range (IQR), the fat horizontal line within the box indicates the median. Whiskers extend to the lowest and highest data points still within 1.5 IQR of the box. Outliers beyond these points are represented by dots.

Figure 3. The Relative Bias in Estimated Regression Coefficients and the Predictive Performance in Relation to the Number of Events per Variable (EPV) and Sample Characteristics (ICC=20%)



Legend: *J*: the number of clusters, n_j: the average number of observations per cluster; *N* (x100): the total number of observations, to be multiplied by 100; *p*: prevalence. A) the relative bias (%) in the estimated regression coefficient $\hat{\beta}_4$; B) the relative within-cluster discrimination (%), computed as $100 \times (C_{within}/C_{sp\ within})$, where $C_{sp\ within}$ is the within-cluster C-index of a model fitted and evaluated in the source population; C) the relative within-cluster calibration slope (%), defined as $100 \times (b_{within}/b_{sp\ within})$, where $b_{sp\ within}$ is the within-cluster calibration slope of a model fitted and evaluated in the source population. The box indicates the interquartile range (IQR), the fat horizontal line within the box indicates the median. Whiskers extend to the lowest and highest data points still within 1.5 IQR of the box. Outliers beyond these points are represented by dots.

Web Material

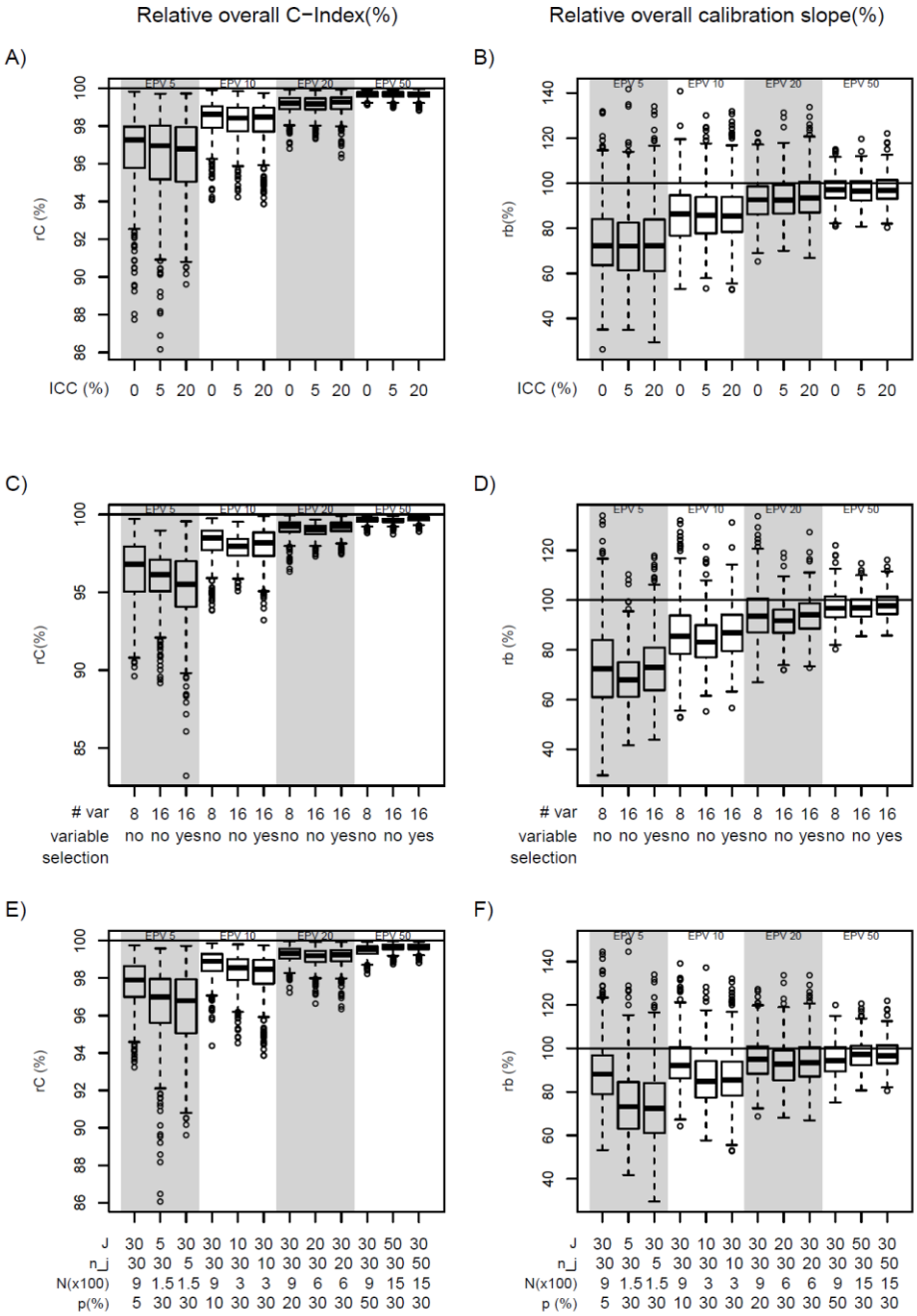
Web Appendix 1. Design of the Simulation Study

Web Appendix 2. Examples of R Code for Simulating Source Populations, Sampling and Model Building

Table A.1. The Relative Bias (%) in the Estimated Regression Coefficients. Bias (Interquartile Range)

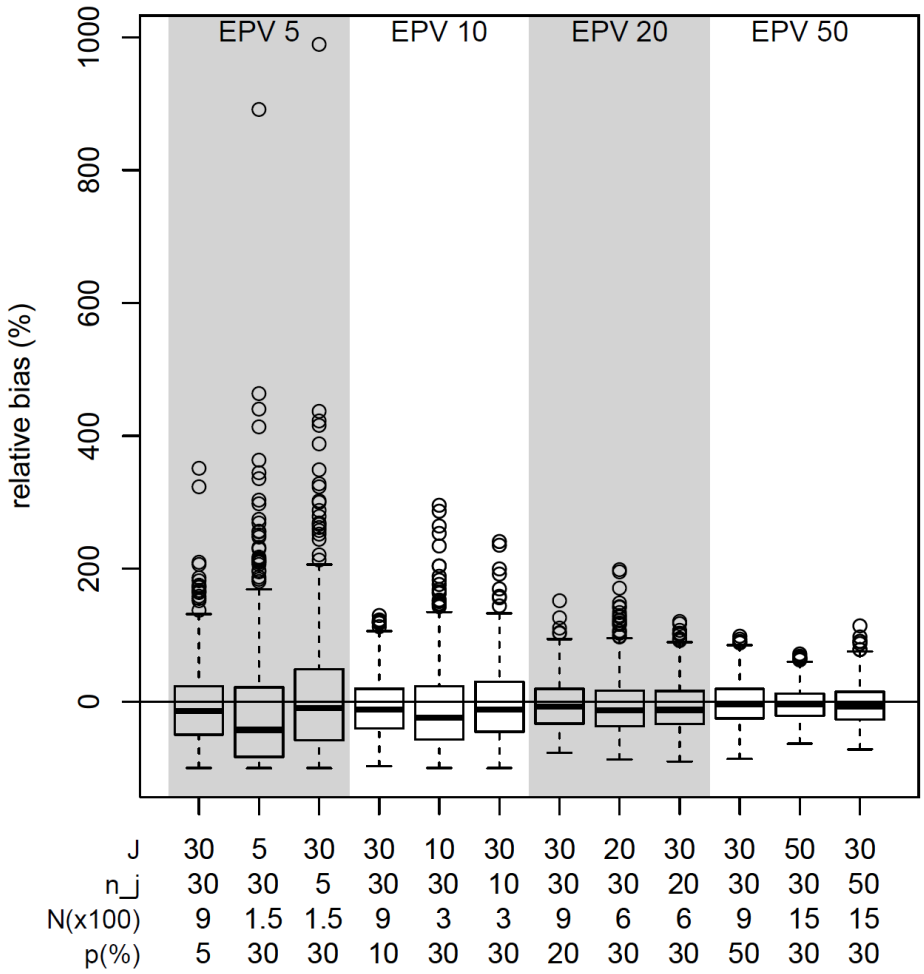
		β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
EPV5	ICC0	7.4 (-13.0 to 33.7)	11.7 (-48 to 63.4)	12.4 (-47.7 to 64.4)	15.0 (-71.2 to 101.5)	6.9 (-39.8 to 59.6)	8.6 (-35.0 to 48.5)	9.0 (-33.5 to 49.1)	9.1 (-28.8 to 51.8)
	ICC5	11.6 (-13.8 to 33.8)	11.5 (-39.2 to 68.2)	14.6 (-37.5 to 72.8)	6.5 (-78.6 to 117.7)	4.1 (-47.6 to 51.0)	5.9 (-33.1 to 52.6)	7.0 (-32.7 to 46.0)	10.6 (-31.8 to 46.3)
	ICC20	5.8 (-15.0 to 32.8)	6.1 (-45.1 to 60.9)	3.2 (-46.6 to 56.5)	-10.2 (-115.1 to 108.9)	5.4 (-48.5 to 54.6)	6.3 (-37.5 to 55.4)	9.5 (-35.9 to 56.6)	5.4 (-31.2 to 49.1)
EPV10	ICC0	5.3 (-7.6 to 19.5)	2.8 (-30.9 to 37.5)	3.5 (-30.4 to 38.4)	-3.1 (-57.3 to 55.5)	4.0 (-28.3 to 32.2)	0.7 (-21.2 to 33.0)	3.8 (-18.5 to 31.1)	4.1 (-22.5 to 29.2)
	ICC5	4.9 (-10.7 to 19.0)	2.0 (-28.3 to 32.9)	4.7 (-26.4 to 36.5)	5.0 (-56.4 to 61.6)	1.0 (-29.8 to 32.5)	2.6 (-21.7 to 28.7)	2.7 (-23.9 to 32.6)	2.5 (-23.4 to 28.9)
	ICC20	4.0 (-9.8 to 19.5)	5.0 (-29.9 to 40.5)	2.1 (-31.9 to 36.7)	8.9 (-60.5 to 74.2)	4.7 (-32.4 to 37.8)	-0.2 (-26.3 to 29.4)	-3.0 (-30.8 to 27.9)	5.1 (-23.3 to 32.3)
EPV20	ICC0	3.5 (-34.5 tot 13.5)	2.2 (-17.4 to 25.3)	2.9 (-16.9 to 26.1)	2.9 (-36.0 to 42.9)	3.1 (-21.9 to 24.3)	3.2 (-14.6 to 20.3)	-0.6 (-19.7 to 19.6)	0.2 (-19.0 to 16.5)
	ICC5	2.7 (-7.7 to 11.7)	2.5 (-23.0 to 21.5)	5.3 (-20.9 to 24.8)	-0.06 (-41.7 to 39.7)	2.4 (-19.8 to 24.3)	-2.6 (-20.6 to 16.8)	2.2 (-16.1 to 19.3)	1.5 (-17.2 to 15.5)
	ICC20	0.7 (-8.7 to 9.9)	5.5 (-15.4 to 29.1)	2.6 (-17.7 to 25.5)	-4.6 (-55.8 to 42.4)	-3.0 (-23.6 to 20.7)	2.7 (-19.2 to 21.3)	4.0 (-17.3 to 23.0)	-1.3 (-18.7 to 20.3)
EPV50	ICC0	1.1 (-4.4 to 7.0)	1.1 (-13.5 to 14.7)	1.7 (-13.0 to 15.5)	-0.4 (-27.1 to 25.1)	0.8 (-11.4 to 11.2)	0.4 (-11.3 to 12.4)	-0.4 (-11.1 to 11.7)	0.7 (-10.3 to 10.9)
	ICC5	1.2 (-5.3 to 7.0)	-1.9 (-14.1 to 11.9)	0.7 (-11.7 to 14.9)	0.9 (-25.4 to 30.9)	1.3 (-10.6 to 14.3)	-0.8 (-10.5 to 11.4)	0.4 (-11.2 to 11.8)	1.4 (-9.9 to 12.3)
	ICC20	0.9 (-5.6 to 7.5)	4.2 (-9.3 to 17.7)	1.3 (-11.8 to 14.5)	-2.1 (-26.1 to 24.7)	-0.6 (-13.6 to 15.0)	0.6 (-12.8 to 11.7)	0.8 (-11.0 to 13.1)	0.0 (-11.2 to 13.2)

Figure A.1. The Relative Overall C-index and the Relative Overall Calibration Slope in Relation to the Number of Events per Variable (EPV), the Amount of Clustering (ICC), Backward Variable Selection and Sample Size Characteristics



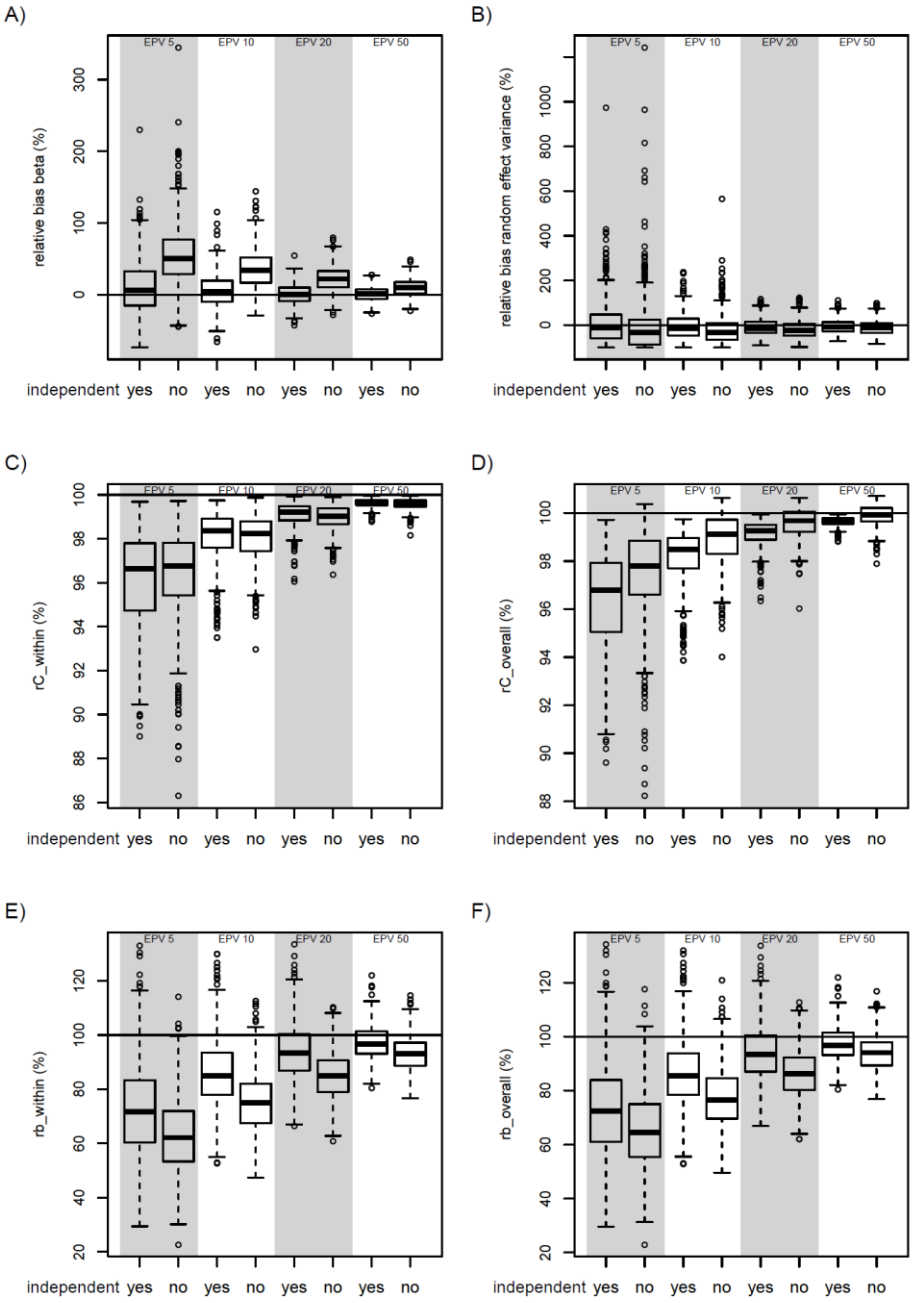
Legend: *rC*: relative overall C-index; *rb*: relative overall calibration slope; #var: number of candidate predictors; *J*: the number of clusters, n_j: the average number of observations per cluster; *N* (x100): the total number of observations, to be multiplied by 100; *p*: prevalence. A) the effect of EPV and ICC on the relative discrimination, defined as $100 \times (C_{overall}/C_{sp\,overall})$, where $C_{sp\ overall}$ is the C-index of a model fitted and evaluated in the source population; B) the effect of EPV and ICC on the relative calibration, defined as $100 \times (b_{overall}/b_{sp\ overall})$, where $b_{sp\ overall}$ is the calibration slope of a model fitted and evaluated in the source population; C) the effect of EPV and backward variable selection on the relative discrimination (ICC=20%); D) the effect of EPV and backward variable selection on the relative calibration (ICC=20%); E) the effect of EPV and sample characteristics on the relative discrimination (ICC=20%); F) the effect of EPV and sample characteristics on the relative calibration (ICC=20%). The box indicates the interquartile range (IQR), the fat horizontal line within the box indicates the median. Whiskers extend to the lowest and highest data points still within 1.5 IQR of the box. Outliers beyond these points are represented by dots.

Figure A.2. The Estimation of the Random Intercept Variance (ICC=20%) in Relation to the Number of Events per Variable (EPV) and Sample Characteristics.



Legend: ICC: intraclass correlation, *J*: the number of clusters, n_j: the average number of observations per cluster; *N* (x100): the total number of observations, to be multiplied by 100; *p*: prevalence. The box indicates the interquartile range (IQR), the fat horizontal line within the box indicates the median. Whiskers extend to the lowest and highest data points still within 1.5 IQR of the box. Outliers beyond these points are represented by dots.

Figure A.3. The Relative Bias in the Estimated Regression Coefficients and the Predictive Performance in Relation to the Number of Events per Variable (EPV) when Predictors are Dependent versus Independent of the Random Intercept (ICC=20%).



ICC: intraclass correlation; rC_{within}: relative within-cluster C-index; $rC_{overall}$: overall relative C-index; rb_{within}: relative within-cluster calibration slope; $rb_{overall}$: relative calibration slope (%). A) relative bias in the estimated regression coefficient $\hat{\beta}_1$; B) estimated random intercept variance; C) relative within-cluster discrimination (%), defined as $100 \times (C_{within}/C_{sp\ within})$, where $C_{sp\ within}$ is the within-cluster C-index of a model fitted and evaluated in the source population; D) relative overall discrimination (%), defined as $100 \times (C_{overall}/C_{sp\ overall})$,, where $C_{sp\ overall}$ is the C-index of a model fitted and evaluated in the source population; E) relative within-cluster calibration (%), computed as $100 \times (b_{within}/b_{sp\ within})$,, where $b_{sp\ within}$ is computed as the within-cluster calibration slope of a model fitted and evaluated in the source population; F) relative overall calibration, defined as $100 \times (b_{overall}/b_{sp\ overall})$,, where $b_{sp\ overall}$ is the calibration slope of a model fitted and evaluated in the source population. The box indicates the interquartile range (IQR), the fat horizontal line within the box indicates the median. Whiskers extend to the lowest and highest data points still within 1.5 IQR of the box. Outliers beyond these points are represented by dots.

Web Appendix 1. Design of the simulation study

This Web Appendix first discusses the multilevel logistic regression model. Second, it presents a stepwise discussion of the generation of source populations for the simulation study. Finally, it gives a stepwise discussion of sampling from the source population and model building.

We studied a 2-level logistic regression model [12] for the probability *p* of an event *Y* occurring for observation *i* in cluster *j*.

$$Y_{ij}|u_j \sim \text{Bernoulli}(p_{ij})$$

$$\text{logit}_{ij} = \log(p_{ij}/(1-p_{ij}))$$

$$\text{logit}_{ij} = \beta_0 + u_j + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8$$

$$u_j \sim \text{N}(0, \sigma_u^2)$$

(A1)

In this equation, X_1 through X_4 are continuous level-1 variables with normal distributions, each with mean 0 and standard deviations 1, 0.6, 0.4, and 0.2, respectively. X_5 through X_8 are level-1 dummy variables with prevalence 0.2, 0.3, 0.3, and 0.4, respectively. All true regression coefficients β are fixed at 0.8. The overall intercept β_0 is set equal to -2.1, to obtain an event rate of the outcome Y_{ij} of 0.3. The random intercept u_j is a normally distributed level-2 term with mean 0 and variance σ_μ^2. Because the logit link function is used, the intraclass correlation (ICC) is equal to $\sigma_\mu^2/(\sigma_\mu^2 + \pi^2/3)$. It can be interpreted as the proportion of the variance in the outcome Y_{ij} that is attributable to clustering. Additionally, we study a 2-level logistic regression model including eight noise predictors. X_9 through X_{16} have the same distributions as X_1 through

X_8, but their true regression coefficients β equal 0. Note that we do not consider any level-2 predictors, i.e. cluster characteristics, or random slopes.

The generation of the source populations was done as follows:

1. The number of clusters (J) was fixed at 200. The number of observations per cluster was drawn from a Poisson distribution with a separate, randomly generated lambda for each cluster. This yielded cluster sizes ranging from approximately 200 to 1000. Hence, each population consisted of approximately 10,000 observations in total.
2. Each cluster was assigned a random cluster intercept u_j. They were generated from a normal distribution of which the standard deviation was determined by the predefined ICC (0%, 5% or 20%) through equation (A2) as follows:

   $\sigma_\mu^2 = (\pi^2/3 \times \text{ICC})/(1 - \text{ICC})$.
3. Each of the Xs was generated from a normal or binomial distribution as defined above.
4. In one source populations (defined as set 6 in Table 1), u_j was correlated with X_1 and X_5. The correlation between X_1 and u_j was induced as follows:

   $X_{1\,correlated\,ij} = \sqrt{0.25} \times u_j + \sqrt{0.75} \times X_{1ij}$, after which $X_{1\,correlated}$ was rescaled to have the same mean and standard deviation as X_1. This yields a correlation of approximately 0.5 between X_1 and u_j.To create $X_{5\,correlated\,ij}$, the success probability determining its distribution was related to u_j:$X_{5\,correlated\,ij} \sim \text{Bernoulli}(0.20 + rescaled\_u_j)$, where $rescaled\_u_j = 0.1 * (u_j - \bar{u_j})/\max[\text{abs}(u_j - \bar{u_j})]$, such that $-0.1 \leq rescaled\_u_j \leq 0.1$ and $rescaled\_u_j \propto u_j$. The prevalence of X_5 hence increased with the random cluster intercept: it is 12% at the cluster with the lowest u_j and 29% at the cluster with the highest u_j.

5. For each of the observations, the logit is computed from the generated *X*s and u_j, using equation (A1). Each β equals 0.8 and the intercept equals -2.1, to obtain a prevalence of Y_{ij} of 0.3.
6. The predicted probabilities of an event occurring are computed by exploiting the inverse logit transformation $p_{ij} = \exp(\text{logit}_{ij})/(1 + \exp(\text{logit}_{ij}))$.
7. Finally, to incorporate an element of randomness in the outcomes, the Y_{ij} are obtained by comparing the predicted probability of the event to a randomly drawn value from a uniform distribution as follows:

$$Z_{ij} \sim \text{unif}(0,1)$$

$$Y_{ij} = \begin{cases} 1 & \text{if } z_{ij} \leq p_{ij} \\ 0 & \text{if } z_{ij} > p_{ij} \end{cases}.$$

(A2)

The sampling from the source population and the model building occurred as follows:

1. The number of clusters (*J*) to be drawn was prespecified according to the condition (see Table 1).
2. The number of events to be sampled was computed from the required EPV and the number of parameters (*k*) to be estimated (17 in sets 2 and 3 and 9 in all other conditions). The total sample size was computed from the number of events and the prevalence in the source population (0.3), or the predefined sample event rates *p* in set 5 (0.05, 0.1, 0.2 or 0.5).
3. Clusters were randomly drawn without replacement from the source population.
4. A list was generated with all observations from the sampled clusters. From this list, events and non-events were sampled separately, hence stratified for outcome Y_{ij}, and

without replacement, to ensure the correct number of events was sampled and the prevalence of the outcome in the sample was as required.

5. In the obtained sample, a random intercept model was fitted. The model either contained the eight predefined predictors (X_1 through X_8), the eight predefined predictors and eight noise variables (X_1 through X_{16}, in set 2), or was built using backward selection (in set 3; α for deletion 0.10). A random intercept model was fitted in each subsequent step of the backward variable selection algorithm, to acknowledge clustering during model building.
6. The fitted model was evaluated in the source population, to obtain the concordance probability, the within-cluster concordance probability, the calibration slope and the within-cluster calibration slope. Predicted probabilities were obtained by omitting the random intercepts from the prediction equation [17].
7. Steps 1 through 6 were repeated 500 times.

Web Appendix 2. Examples of R code for Simulating Source Populations, Sampling and Model Building

1. **Generation of source populations**

```
rm(list=ls(all=TRUE))
library (rms)
library (lme4)

#Step 1. Fix the number of clusters (nb) and draw the number of observations per cluster (n2)
from a poisson distribution.
nb <- 200
mean <- 6.22
sd <- 0.3
L1<-rnorm(nb, mean, sd)
L<-round(exp(L1))
n2 <- rpois(nb, L)
an <-seq(1:nb)
anest <-rep(an, n2)
n <- length(anest) #number of observations in the source population
pat_nr <- seq(1:n)

#Step 2. Generate random cluster intercepts (randeff).
randeff <- rnorm(nb, 0, 0.416)  # this is to obtain an ICC of 5%
```

```
randeffect <-rep(randeff, n2)


## Step 3. Generate predictors x1 to x8
x1      <- rnorm(n, 0, 1)
x2      <- rnorm(n, 0, 0.6)
x3      <- rnorm(n, 0, 0.4)
x4      <- rnorm(n, 0, 0.2)
x5      <- rbinom(n, 1, 0.20)
x6      <- rbinom(n, 1, 0.30)
x7      <- rbinom(n, 1, 0.30)
x8      <- rbinom(n, 1, 0.40)


#Step 4: Introduce correlation between random intercept and predictors if required.
# Step 5: Set beta coefficients and obtain the logit.
b1      <- 0.8
b2      <- 0.8
b3      <- 0.8
b4      <- 0.8
b5      <- 0.8
b6      <- 0.8
b7      <- 0.8
b8      <- 0.8
logit   <- -2.1+b1*x1+b2*x2+b3*x3+b4*x4+b5*x5+b6*x6+b7*x7+b8*x8+randeffect
```

```
#Step 6: Calculate the predicted probabilities of experiencing an event
p       <- plogis(logit)

#Step 7: Obtain dichotomous outcome y
y       <- ifelse(runif(n)<=p, 1, 0)
data <- as.data.frame(cbind(pat_nr, x1, x2, x3, x4, x5, x6, x7, x8, randeffect, anest, y), dimnames
= list(c(1:n), Cs(pat_nr, x1, x2, x3, x4, x5, x6, x7, x8, randeffect, anest, y)))
```

2. Sampling from the source population and model building within samples

This code corresponds to simulation condition 1.11 (Table 1). The code can be adjusted for other conditions with minor changes.

```
rm(list=ls(all=TRUE))
library (Hmisc)
library (rms)
library (lme4)
library(arm)
setwd("C:\\path for source population data")
load('name of source population') #load the source population
set.seed(choose a seed)
Nsample <- 500  #specify the number of samples to be drawn

#Step 1. Specify the number of clusters
```

```
Nsamplev2 <-30

#Step 2. Specify the number of events to be drawn
NEVENTS <- 450
NNONEVENTS <- round((NEVENTS/mean(data$y))-NEVENTS)

#Matrices to store results
Nanest_sample <- matrix(NA, nrow=Nsample, ncol=1)
anest_sample <- matrix(NA, nrow=Nsample, ncol=Nsamplev2)
Nevents_sample <- matrix(NA, nrow=Nsample, ncol=1)
Npats_sample <- matrix(NA, nrow=Nsample, ncol=1)
Pevents_sample <- matrix(NA, nrow=Nsample, ncol=1)
Npats_anest_sample <- matrix(NA, nrow=Nsample, ncol=Nsamplev2)

mat_sd_ranef0 <- matrix(NA, nrow=Nsample, ncol=1)
mat_sd_ranef1 <- matrix(NA, nrow=Nsample, ncol=1)
mat_regcof <- matrix(NA, nrow=Nsample, ncol=9)
mat_SEregcof <- matrix(NA, nrow=Nsample, ncol=9)
mat_ranef <- matrix(NA, nrow=Nsample, ncol=Nsamplev2)
mat_SEranef <- matrix(NA, nrow=Nsample, ncol=Nsamplev2)
Niterm1 <- matrix(NA, nrow=Nsample, ncol=1)
Equaldeviance <- matrix(NA, nrow=Nsample, ncol=1)
mat_dims <- matrix(NA, nrow=Nsample, ncol=18)
```

```
realvalresultB <- matrix(NA,nrow=Nsample, ncol=5)
dimnames(realvalresultB) <- list(c(1:Nsample),
Cs(C,calib_slope,Cw,ML_calib_slope,var_ML_calib_slope))

loglik0 <- matrix(NA,nrow=Nsample,ncol=1)
loglik1 <- matrix(NA,nrow=Nsample,ncol=1)

#Start sampling from the domain
tmpA <- aggregate(data$pat_nr, list(anest = data$anest), FUN=length)
pt_anest <- as.vector(tmpA[,2])
names(pt_anest) <- tmpA[,1]

for (r in 1:Nsample){

#Step 3. Sample clusters (samp_anest) from the source population
samp_anest <- sample(x=names(pt_anest), size=Nsamplev2, replace = F)

#Get the data from sampled clusters (dsample_lev2)
loc.pt_anest <- pt_anest[samp_anest]
indices.cutoffs <- c(0,cumsum(loc.pt_anest))
ptnrs_samp2 <- matrix(NA,nrow=sum(loc.pt_anest),ncol=1)
for (i in 1:Nsamplev2){
```

```
ptnrs_samp2[(indices.cutoffs[i]+1):indices.cutoffs[i+1],] <-
data[data$anest==names(loc.pt_anest)[i],1]
}
dsample_lev2 <- data[ptnrs_samp2,]

#Step 4. Stratified sampling of events (sample_lev1_y1) and non-events (sample_lev1_y0).
sdatay1 <- dsample_lev2[dsample_lev2$y==1,]
pt_nr_y1 <- as.vector(sdatay1$pat_nr)
sample_lev1_y1 <- sample(pt_nr_y1, NEVENTS , replace = FALSE)
sdatay0 <- dsample_lev2[dsample_lev2$y==0,]
pt_nr_y0 <- as.vector(sdatay0$pat_nr)
sample_lev1_y0 <- sample(pt_nr_y0, NNONEVENTS, replace = FALSE)
sample_lev1 <- c(sample_lev1_y0, sample_lev1_y1)
dsample <- data[sample_lev1,]

#Step 5. Fit a random intercept model in the sample: full model (m1) and null model (m0)
iter_m1<-capture.output(m1<-lmer(y~x1+x2+x3+x4+x5+x6+x7+x8+(1|anest), family =
"binomial", data=dsample, verbose=T ))
m0<-lmer(y~(1|anest), family = "binomial", data=dsample)
Last2it<-iter_m1[c(length(iter_m1)-1):c(length(iter_m1))]
Last2itsplit<-strsplit(Last2it,":")
Equal<- Last2itsplit [[1]][2]== Last2itsplit [[2]][2]#check convergence
```

```
# Set Npats_anest, anest and ranef to missing when we sampled <1 observation per cluster
anest_s <- length(unique(dsample$anest))
N_missing_anest <- Nsamplev2-Nanest_s
add_anest <- rep(x=-99, times=N_missing_anest)
if (Nanest_s < Nsamplev2){
anest_s <- c(unique(dsample$anest), add_anest)
Npats_anest_s <- c((aggregate(dsample$pat_nr, list(anest = dsample$anest),
FUN=length)[,2]),add_anest)
clustereff_s <- c(attr(m1, "ranef"),add_anest)
clustereff_s_SE <- c(as.numeric(se.ranef(m1)$anest), add_anest)
}
if (Nanest_s == Nsamplev2) {
Npats_anest_s <- aggregate(dsample$pat_nr, list(anest = dsample$anest), FUN=length)[,2]
anest_s <- unique(dsample$anest)
clustereff_s <- attr(m1, "ranef")
clustereff_s_SE <- as.numeric(se.ranef(m1)$anest)
}

#save sample summaries
Nanest_sample[r,] <- Nanest_s
anest_sample[r,] <- anest_s
Npats_anest_sample[r,] <- Npats_anest_s
Nevents_sample[r,] <- sum(dsample$y)
```

```
Npats_sample[r,] <- length(dsample[,1])

Pevents_sample[r,] <- mean(dsample$y)


#save model parameters

mat_sd_ranef0[r,]<- sigma.hat(m0)$sigma$anest

mat_sd_ranef1[r,]<- sigma.hat(m1)$sigma$anest

mat_regcof[r,]<- as.numeric(fixef(m1))

mat_SEregcof[r,]<- se.coef(m1)$fixef

mat_ranef[r,] <- clustereff_s

mat_SEranef[r,] <- clustereff_s_SE

mat_dims[r,] <- as.numeric(attr(m1,"dims"))

Niterm1[r,] <- length(iter_m1)-1

Equaldeviance[r,] <- Equal


#save log likelihood of full and null model

loglik0[r,]<-logLik(m0)

loglik1[r,]<-logLik(m1)


#Step 6. Evaluate model performance in the source population

data$lpB <- cbind(1, as.matrix(data[,2:9])) %*% fixef(m1)

perfm_m1B<- lrm(data$y~data$lpB)

realvalresultB[r,1] <- perfm_m1B$stats[6] # C

realvalresultB[r,2] <- perfm_m1B$coefficients[2] #calibration slope
```

```
C.w.index <- rep(NA, length(unique(data$anest))) #within-cluster C. Requires function
concordance.prob.logistic from Van Oirbeek et al 2012.
n.comp <- rep(NA, length(unique(data$anest)))
for(p in 1:length(unique(data$anest))){
data_p <- data[data$anest==p,]
prediction <- plogis(data_p$lpB)
outcome <- data_p$y
result <- concordance.prob.logistic(outcome, prediction)
C.w.index[p] <- result$C
n.comp[p] <- result$n.comp
}
realvalresultB[r,3] <- sum(C.w.index*n.comp/sum(n.comp, na.rm = TRUE), na.rm = TRUE)
#within-cluster C-index
calibslope_mlB <- glmer(y~lpB+(lpB|anest), family = "binomial", data=data)
realvalresultB[r,4] <- as.numeric(fixef(calibslope_mlB)[2]) #within-cluster calibration slope
realvalresultB[r,5] <- (sigma.hat(calibslope_mlB)$sigma$anest[2])^2 #random variance of the
calibration slope

#save to check convergence
write.table(cbind(r, Niterm1[r,], iter_m1), file="C:\\path\\iter", col.names=F, row.names=F,
append=T)
write.table(x=Equaldeviance, file=" C:\\path\\Equaldeviance", append=F, na="NA",
col.names=F)
```

```
}
#Save all results
write.table(x=Nanest_sample, file="C:\\path\\Nanest", append=F, na="NA", col.names=F)
#Number of clusters sampled
write.table(x=anest_sample, file="C:\\path\\anest", append=F, na="NA", col.names=F) # ID
numbers of sampled clusters
write.table(x=Nevents_sample, file="C:\\path\\Nevents", append=F, na="NA", col.names=F)
#number of events sampled
write.table(x=Npats_sample, file="C:\\ path \\Npats", append=F, na="NA", col.names=F)
#Number of observations sampled
write.table(x=Pevents_sample, file="C:\\ path \\Pevents", append=F, na="NA", col.names=F)
#Event rate of the outcome in the sample
write.table(x=Npats_anest_sample, file="C:\\ path \\Npats_anest_sample", append=F, na="NA",
col.names=F) #Sampled number of observations per cluster
write.table(x=mat_sd_ranef0, file="C:\\ path \\sdranef0", append=F, na="NA", col.names=F)
#random effect standard deviation of the null model
write.table(x=mat_sd_ranef1, file="C:\\ path \\sdranef1", append=F, na="NA", col.names=F)
#random effect standard deviation of the model
write.table(x=mat_regcof, file="C:\\ path \\regcof", append=F, na="NA", col.names=F)
#Estimated regression coefficients of the prediction model
write.table(x=mat_SEregcof, file="C:\\ path \\SEregcof", append=F, na="NA", col.names=F)
#standard errors of regression coefficients
```

write.table(x=mat_ranef, file="C:\\ *path* \\ranef", append=F, na="NA", col.names=F) #random cluster intercepts

write.table(x=mat_SEranef, file="C:\\ *path* \\SEranef", append=F, na="NA", col.names=F) #standard errors of random cluster intercepts

write.table(x=loglik1, file="C:\\ *path* \\loglik_1", append=F, na="NA", col.names=F) #log likelihood of the model

write.table(x=loglik0, file="C:\\ *path* \\loglik_0", append=F, na="NA", col.names=F) #log likelihood of the null model

write.table(x=realvalresultB, file="C:\\ *path* \\realvalresultB", append=F, na="NA", col.names=F) #predictive performance measures

References

1. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York, NY: Springer US; 2009.

2. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373-1379. DOI: 10.1016/S0895-4356(96)00236-3.

3. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361-387. DOI: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

4. Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol. 1999;52(10):935-942. DOI: 10.1016/S0895-4356(99)00103-1.

5. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. J Clin Epidemiol. 2011;64(9):993-1000. DOI: 10.1016/j.jclinepi.2010.11.012.

6. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol. 2007;165(6):710-718. DOI: 10.1093/aje/kwk052.

7. Sprague S, Matta JM, Bhandari M, Dodgin D, Clark CR, Kregor P, et al. Multicenter collaboration in observational research: improving generalizability and efficiency. J Bone Joint Surg Am. 2009;91 Suppl 3:80-86. DOI: 10.2106/jbjs.h.01623.

8. Senn S. Some controversies in planning and analysing multi-centre trials. Stat Med. 1998;17(15-16):1753-1765; discussion 1799-1800. DOI: 10.1002/(SICI)1097-0258(19980815/30)17:15/16<1753::AID-SIM977>3.0.CO;2-X.

9. Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for center in multicenter studies: an overview. Ann Intern Med. 2001;135(2):112-123. DOI: 10.7326/0003-4819-135-2-200107170-00012.

10. Bouwmeester W, Twisk J, Kappen T, Klei W, Moons K, Vergouwe Y. Prediction models for clustered data: comparison of a random intercept and standard regression model. BMC Med Res Methodol. 2013;13(1). DOI: 10.1186/1471-2288-13-19.

11. Debray TPA, Moons KGM, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat Med. 2013;32(18):3158-3180. DOI: 10.1002/sim.5732.

12. Snijders TAB, Bosker RJ. Multilevel analysis : an introduction to basic and advanced multilevel modeling. 2nd ed. Snijders TAB, Bosker RJ, editors. London: Sage; 2012.

13. Maas CJM, Hox JJ. Sufficient Sample Sizes for Multilevel Modeling. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences. 2005;1(3):86-92. DOI: 10.1027/1614-2241.1.3.86

14. Moineddin R, Matheson FI, Glazier RH. A simulation study of sample size for multilevel logistic regression models. BMC Med Res Methodol. 2007;7(34). DOI: 10.1186/1471-2288-7-34.

15. Paccagnella O. Sample Size and Accuracy of Estimates in Multilevel Models. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences. 2011;7(3):111-120. DOI: 10.1027/1614-2241/a000029.

16. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. J Clin Epidemiol. 2004;57(8):785-794. DOI: 10.1016/j.jclinepi.2003.12.013.

17. Skrondal A, Rabe-Hesketh S. Prediction in multilevel generalized linear models. Journal of the Royal Statistical Society Series A (Statistics in Society). 2009;172(3):659-687. DOI: 10.1111/j.1467-985X.2009.00587.x.

18. Harrell FE. Regression modeling strategies : with applications to linear models, logistic regression, and survival analysis. Harrell FE, Jr., editor. New York, NY: Springer; 2001.

19. Cox DR. Two Further Applications of a Model for Binary Regression. Biometrika. 1958;45(3/4):562-565. DOI: 10.2307/2333203.

20. Van Oirbeek R, Lesaffre E. Assessing the predictive ability of a multilevel binary regression model. Computational Statistics & Data Analysis. 2012;56(6):1966-1980. DOI: 10.1016/j.csda.2011.11.023.

21. R Development Core Team. R: A language and environment for statistical computing Vienna, Austria: R Foundation for Statistical Computing; 2011 [October 7, 2014]. Available from: http://www.R-project.org/.

22. Bates D, Maechler M, Bolker B. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-42. 2011 [October 7, 2014]. Available from: http://CRAN.R-project.org/package=lme4.

23. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med. 2000;19(8):1059-1079. DOI: 10.1002/(SICI)1097-0258(20000430)19:8<1059::AID-SIM412>3.0.CO;2-0.

24. Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) group. Ultrasound in Obstetrics and Gynecology. 2000;16(5):500-505. DOI: 10.1046/j.1469-0705.2000.00287.x.

25. Kaijser J, Bourne T, Valentin L, Sayasneh A, Van Holsbeke C, Vergote I, et al. Improving strategies for diagnosing ovarian cancer: a summary of the International Ovarian Tumor Analysis (IOTA) studies. Ultrasound Obstet Gynecol. 2013;41(1):9. DOI: 10.1002/uog.12323.

26. van Klaveren D, Steyerberg E, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. BMC Med Res Methodol. 2014;14(1). DOI: 10.1186/1471-2288-14-5.